

Plan général :

- 0 - Préliminaires : rappels de probabilité (1 séance d'exercices)
- 1 - Lois multidimensionnelles (≈ 5 séances)
- 2 - Estimation (≈ 6 séances)
- 3 - Tests (≈ 4 séances)
- 4 - Statistique bayésienne (≈ 4 séances).

+ 1 séance de retour sur l'examen mi-parcours, soit 21 séances au global.

Organisation: Cours partagé entre Fabienne CASTELL et Xavier MILHAUD.

Chaque séance de 3h comporte

1h30 de cours (CM)	{
1h30 de TD	

L'évaluation est réalisée à partir de TP notés, d'un partielle (examen de mi-parcours) et d'un examen terminal. Il y a 4 séances de TP de 4h, dont 1 ou 2 sont notées.

CHAP 2 : Estimation

Avant-propos: cette partie de cours est largement inspirée des cours d'olivier GAUDIEN, reçus à l'ENSIMAG.

I. Introduction

→ L'objectif de la statistique descriptive est de décrire et de résumer l'information contenue dans les données avec des indicateurs statistiques simples (moyenne, écart-type, quantile, ...) et des graphiques (bâtons, histogramme, ...).

→ L'objectif de la statistique inférentielle est de pouvoir faire des prévisions et prendre des décisions au vu des données observées. Il existe 2 grandes catégories de techniques pour le faire :

- l'estimation, ponctuelle ou ensembliste (intervalle de confiance), avec principalement la méthode des moments et celle du maximum de vraisemblance;
- les tests (d'hypothèse) pour la prise de décision.

→ Il existe également la statistique bayésienne, dont le but est d'intégrer l'expérience des données observées pour la mise-à-jour d'une connaissance à priori, représentée par une loi de probabilité modélisant l'incertitude sur le paramètre de la loi modélisant le phénomène d'intérêt.

On distingue en général la statistique paramétrique qui suppose l'existence d'un modèle connu avec des paramètres inconnus (θ inférés), de la statistique non-paramétrique, qui ne fait pas ce type d'hypothèse.

→ Lien entre statistique et théorie des probabilités via le modèle statistique. Un modèle statistique est un objet mathématique associé à l'observation de données issues d'un phénomène aléatoire.

Une expérience statistique recueille une observation x d'un aléa X , à valeurs dans un espace \mathcal{X} et dont on ne connaît pas exactement la loi de probabilité P . On admet ensuite que P appartient à une famille \mathcal{P} de lois de probabilités possibles.

Définition : Le modèle statistique associé à cette expérience est le triplet (X, \mathcal{A}, P) où :

- X est l'espace des observations, ensemble de toutes les observ. possibles.
- \mathcal{A} est la tribu des événements observables associée.
- P est une famille de lois de probabilités possibles définies sur \mathcal{A} .

- Le modèle est discret si X est fini ou dénombrable. Alors la tribu \mathcal{A} est l'ensemble des parties de X : $\mathcal{A} = P(X)$. C'est le cas quand l'événement X observé a une loi de probabilité discrète.
- Le modèle est continu quand $X \subset \mathbb{R}^P$ et $\forall P \in \mathcal{P}$, P admet une densité par rapport à la mesure de Lebesgue dans \mathbb{R}^P . Ici, \mathcal{A} est la tribu des boréliens de X (engendré par les ouverts de X): $\mathcal{A} = \mathcal{B}(X)$.
- Il peut arriver que l'observation de l'événement X ait certains éléments discrets et d'autres continus: X et \mathcal{A} sont alors plus complexes.

Le plus souvent, l'élement aléatoire observé est constitué de variables aléatoires indépendantes et de même loi (iid). On note $X = (X_1, \dots, X_n)$ où les X_i sont iid. On parle alors d'un modèle d'échantillon.

Dans ce cas, en notant (X, \mathcal{A}, P) le modèle statistique d'un échantillon de taille 1, on notera $(X, \mathcal{A}, P)^n$ le modèle correspondant à un échantillon de taille n .

→ Exemples:

- 1) Durée de vie de pièces mécaniques en auto. On recueille n durées de vie de ces pièces, supposées \perp et de même loi exponentielle. On a donc un modèle statistique de la forme $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+), \{\exp(\lambda); \lambda \in \mathbb{R}^+\})^n$.
- 2) Qualité de certaines pièces: on s'intéresse à la proportion de pièces défectueuses. L'observation est donc $x = (x_1, x_2, \dots, x_n)$ où $x_i = \begin{cases} 0 & \text{si OK} \\ 1 & \text{si défectueux} \end{cases}$. On admet le modèle statistique $(\{0, 1\}, \mathcal{D}_{\{0, 1\}}, \{B(p); p \in [0, 1]\})^n$ Bernoulli.

→ Retour sur les modèles paramétrique et non-paramétrique.

- Un modèle paramétrique est un modèle où l'on suppose que le type de loi de X est connu, mais qu'il dépend d'un paramètre θ inconnu (de dimension $d \geq 1$). Alors la famille de lois de probabilité possibles pour X peut s'écrire $\mathcal{P} = \{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\}$. (cas des 2 exemples ci-dessus). Dans ce contexte, on cherche à faire de l'inference statistique sur θ (estimation, test): on dit qu'on fait de la statistique paramétrique.
- Un modèle non-paramétrique est un modèle où \mathcal{P} ne peut pas s'écrire sous cette forme. Dans ce cadre, \mathcal{P} peut par exemple être l'ensemble des lois de probabilités continues sur \mathbb{R} , ou l'ensemble des lois de probabilités sur \mathbb{R} symétriques par rapport à l'origine, ... Les objets sur lesquels portent les procédures d'estimation et de test ne sont plus des paramètres de lois de probabilité. On peut vouloir estimer des quantités comme l'espérance et la variance (avec des estimateurs empiriques par exemple), faire un test sur la valeur d'une espérance, l'adéquation à une loi... C'est de la statistique non-paramétrique.

Un des problèmes de la statistique paramétrique est l'erreur due à un mauvais choix de modèle. L'avantage de la statistique non paramétrique est de ne pas être soumise à ce risque; par contre, si les observations sont bien issues d'un modèle précis, les méthodes statistiques paramétriques utilisant ce modèle seront plus performantes que celles ne l'utilisant pas.

II - Estimation ponctuelle

① - Introduction

On suppose dans toute cette partie que nous faisons de la statistique paramétrique dans un cadre unidimensionnel ($d=1$). De plus, on suppose que les données x_1, \dots, x_n sont n réalisations indépendantes d'une même v.a. X , donc nous sommes dans un modèle d'échantillon. Il est équivalent de supposer que les observations x_1, \dots, x_n sont des réalisations de v.a. X_1, \dots, X_n iid.

- La statistique descriptive permet de faire des hypothèses sur le type de loi de probabilité des X_i .
- Des techniques plus sophistiquées (comme les tests d'adéquation) permettent de valider ou non ces hypothèses.
- ⇒ Ici, on suppose que ces techniques ont permis de choisir une famille de lois de probabilité bien précise pour la loi des X_i , avec un paramètre inconnu. On note θ ce paramètre que l'on cherchera à estimer.
Au vu des observations x_1, \dots, x_n , on aimerait fournir une approximation la plus proche possible de la vraie valeur inconnue.

Cette approximation peut se faire sous la forme d'une unique valeur (pointuelle) ou un ensemble de valeurs vraisemblables (région de confiance).

Dans la suite, on note $F(x; \theta)$ la fonction de répartition (FDR) des X_i .

Dans le cas discret, on note $P(X=x; \theta)$ les probabilités élémentaires -

Dans le cas continu, on note $f(x; \theta)$ la densité de probabilité.

Exemple : si $X \sim \text{Exp}(\lambda)$ alors $F(x; \lambda) = 1 - e^{-\lambda x}$ et $f(x; \lambda) = \lambda e^{-\lambda x}$.

② - Méthodes d'estimation.

Plusieurs techniques d'estimation existent pour un paramètre θ . On peut par exemple utiliser le graphique de probabilité, ou bien le fait que une probabilité puisse s'estimer par une proportion. Ici, nous présentons les 2 méthodes d'estimation les plus communes : la méthode des moments, et la méthode du maximum de vraisemblance.

a) Définition d'un estimateur :

On observe les données x_1, x_2, \dots, x_n . Pour estimer θ , on ne dispose que de cette information. Ainsi, un estimateur de θ sera une fonction de ces observations.

Définition : Une statistique t est une fonction des observations x_1, \dots, x_n :

$$t : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

$$(x_1, \dots, x_n) \mapsto t(x_1, \dots, x_n).$$

→ Exemple : $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$ est une statistique.

- x_1^* (la 1ère observation lorsqu'on a tiré l'échantillon) est une statistique

$$L = \min(x_1, \dots, x_n)$$

Nous avons vu que x_1, \dots, x_n sont des réalisations des v.a. X_1, \dots, X_n . (4)

Ainsi $t(x_1, \dots, x_n)$ est une réalisation de la v.a. $t(X_1, \dots, X_n)$.

→ Exemple: $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ est une réalisation \bar{x}_n .

→ Notation: On note souvent $t_n = t(x_1, \dots, x_n)$ et $T_n = t(X_1, \dots, X_n)$ pour simplifier les écritures.

Définition: Un estimateur de θ est une statistique T_n (donc une v.a.) à valeurs dans l'ensemble des valeurs possibles de θ . Une estimation de θ est une réalisation t_n de l'estimateur T_n .

b) La méthode des moments

→ Idee de base: estimer une espérance mathématique par une moyenne empirique, estimer une variance par une variance empirique... puis procéder par identification après avoir fait le lien entre le paramètre de la loi et ces quantités.

→ Exemple: si le paramètre à estimer est également l'espérance de la loi des X_i (loi de Poisson par exemple), alors on peut l'estimer par la moyenne empirique de l'échantillon. Plus formellement, si $\theta = E[X]$, alors l'estimateur par la méthode des moments (EMM) vaut $\tilde{\theta}_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

→ Plus généralement, pour $\theta \in \mathbb{R}$, si $E[X] = \varphi(\theta)$ où φ est une fonction inversible, alors l'EMM de θ vaut $\tilde{\theta}_n = \varphi^{-1}(\bar{X}_n)$.

→ Si la loi des X_i a 2 paramètres θ_1 et θ_2 (ex: gaussienne), on utilise alors le moment d'ordre 2: ainsi $(\tilde{\theta}_{1n}, \tilde{\theta}_{2n}) = \varphi^{-1}(\bar{X}_n, S_n^2)$ où $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ est un estimateur de la variance $Var(X)$.

→ Exemples :

- Loi de Bernoulli : $X_i \stackrel{iid}{\sim} \mathcal{B}(p)$, avec $E[X] = p$.
Ainsi $\tilde{p}_n = \bar{X}_n$: donc l'estimateur est la proportion de "1" dans l'échantillon.
- Loi exponentielle : $X_i \sim \text{Exp}(\lambda)$ avec $E[X] = 1/\lambda$.
- Loi normale : $X_i \sim \mathcal{N}(\mu, \sigma^2)$ avec $\begin{cases} E[X] = \mu \\ \text{Var}(X) = \sigma^2 \end{cases}$
- Loi Gamma : $X_i \sim G(a, \lambda)$ avec $\begin{cases} E[X] = a\lambda \\ \text{Var}(X) = a/\lambda^2 \end{cases}$

c) La méthode du maximum de vraisemblance:

→ La fonction de vraisemblance : comme son nom l'indique, elle servira à évaluer à quel point tel ou tel modèle est vraisemblable au vu des données.

Définition : quand les observations sont toutes discrètes ou toutes continues, on appelle fonction de vraisemblance (ou vraisemblance), pour l'échantillon x_1, \dots, x_n la fonction du paramètre θ :

$$L(\theta; x_1, \dots, x_n) = \begin{cases} \prod_{i=1}^n P(X_i = x_i, \dots, X_n = x_n; \theta) & \text{dans le cas discret,} \\ f_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta) & \text{dans le cas continu.} \end{cases}$$

→ En utilisant l'hypothèse iid des observations et des v.s. sous-jacentes, X_i , la vraisemblance s'écrit aussi : (on passe de loi multivariée à univariée).

$$L(\theta; x_1, \dots, x_n) = \begin{cases} \prod_{i=1}^n P(X_i = x_i; \theta) = \prod_{i=1}^n P(X = x_i; \theta) \\ \prod_{i=1}^n f_{X_i}(x_i; \theta) = \prod_{i=1}^n f_X(x_i; \theta). \end{cases}$$

→ La vraisemblance est considérée comme une fonction du paramètre θ , et dépend des observations x_1, \dots, x_n .

→ Intuition par l'exemple: Considérons une unique observation, $n=1$. ⑤

On considère savoir que $X_1 \sim \mathcal{B}(15, p)$ avec p inconnu.

On observe $x_1 = 5$ et on cherche à estimer p . Théoriquement, la

vraisemblance s'écrit: $L(p; 5) = P(X_1=5; p) = C_{15}^5 p^5 (1-p)^{15-5}$.

→ C'est la probabilité d'avoir observé 5 quand le paramètre vaut p .

Faisons donc varier p pour déterminer pour quelle valeur de p cette probabilité est maximale:

p	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
$L(p; 5)$	0,01	0,1	0,21	0,13	0,09	0,02	0,003	10^{-4}	$2 \cdot 10^{-7}$

On constate par exemple que la probabilité d'observer 5 si $p=0,8$ est de $1/10000$. Elle est maximale si $p=0,3$ dans cette grille (ce qui est logique du reste puisqu'il y a 15 expériences de Bernoulli).

Il est donc beaucoup plus vraisemblable que $p=0,3$. On cherche ainsi la valeur de p qui maximise cette fonction de vraisemblance, parmi toutes les valeurs possibles de p .

→ Optimisation (maximisation) d'une fonction: pour mener cette recherche du meilleur paramètre, on maximise la vraisemblance en recherchant le point qui annule sa dérivée. Par ailleurs la vraisemblance est définie comme un produit. Il est plus simple de maximiser une somme qu'un produit ⇒ on maximise la log-vraisemblance. Dans l'exemple ci-dessus, elle vaut:

$$\ln L(p; x_1) = \ln C_{15} + x_1 \ln p + (15-x_1) \ln (1-p)$$

et admet donc comme dérivée $\frac{d}{dp} \ln L(p; x_1) = \frac{x_1}{p} - \frac{15-x_1}{1-p} = \frac{x_1 - 15p}{p(1-p)}$

Ainsi, la dérivée s'annule pour $p = \frac{x_1}{15} = \frac{5}{15} = \frac{1}{3}$.

→ La valeur la + vraisemblable de p est $1/3$, avec comme vraisemblance la vraisemblance maximale $L\left(\frac{1}{3}; 5\right) = 21,4\%$.

Définition : L'estimation de maximum de vraisemblance de θ est la valeur $\hat{\theta}_n$ de θ qui rend maximale la fonction de vraisemblance $L(\theta; x_1, \dots, x_n)$. L'estimateur de maximum de vraisemblance (EMV) de θ est la variable aléatoire correspondante.

→ Ainsi, on exprime $\hat{\theta}_n$ comme $\hat{\theta}_n = \underset{\theta}{\operatorname{argmax}} \ln L(\theta; x_1, \dots, x_n)$

Lorsque $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$ et que toutes les dérivées partielles existent, $\hat{\theta}_n$ est solution du système d'équations appelées équations de vraisemblance.

$$\forall j \in \{1, \dots, d\}, \quad \frac{\partial}{\partial \theta_j} \ln L(\theta; x_1, \dots, x_n) = 0.$$

Rq: Une solution de cette optimisation pourrait a priori être un minimum. Mais la nature même d'une fonction de vraisemblance conduit à un maximum!

• Lorsque le système n'a pas de solution explicite, on utilise une méthode numérique - la plus commune est l'algorithme de Newton-Raphson.

→ Exemples:

• Loi de Bernoulli:

• Loi exponentielle:

• Loi normale:

• Loi Gamma:

→ Monter une optimisation en R? (avec optim).

Rq: En toute généralité, on obtient 2 estimateurs \neq (EMM et EMV) ⑥
 selon la méthode pour estimer le paramètre de la loi. Cela amène
une question naturelle: quel est le meilleur?

Exercice - Annex Chap 6 - 1, 2, 3

(3) Qualité d'un estimateur

On rappelle qu'on s'intéresse ici à un paramètre θ de dimension 1. En particulier, $\theta \in \mathbb{R}$. Les estimateurs T_n sont ici des v.a. réelles.

Pour $\theta \in \mathbb{R}^d$ avec $d \geq 2$, toutes les notions sont généralisées mais sont + complexes. Par exemple, la notion de variance est remplacée par celle de matrice de covariance.

a) Biais et variance d'un estimateur

Un estimateur T_n de θ est un bon estimateur, si il est "proche" de θ . Cela nécessite de définir une mesure de l'écart entre T_n et θ : c'est le risque de l'estimateur, que l'on cherche à minimiser.

→ Exemple de risque: $T_n - \theta$, $(T_n - \theta)^2$, $|T_n - \theta|$.

On s'intéresse en réalité à la version déterministe de ces quantités, obtenues par passage à l'espérance. En particulier, ...

Définition : • Le biais de T_n vaut $E[T_n - \theta] = E[T_n] - \theta$.

• Le risque quadratique (EQM) vaut $EQM(T_n) = E[(T_n - \theta)^2]$.

Définition : Un estimateur T_n de θ est dit sans biais $\Leftrightarrow E[T_n] = \theta$.
 Il est biaisé $\Leftrightarrow E[T_n] \neq \theta$.

Interprétation : Le biais mesure une erreur systématique d'estimation de θ par T_n . Si $E[T_n] - \theta < 0$ (biais négatif) alors T_n a tendance à sous-estimer la vraie valeur θ .

→ On peut réécrire l'EQM :

$$\begin{aligned} EQM(T_n) &= E[(T_n - \theta)^2] = E[(T_n - E[T_n] + E[T_n] - \theta)^2] \\ &= E[(T_n - E[T_n])^2 + 2(T_n - E[T_n])(E[T_n] - \theta) + (E[T_n] - \theta)^2] \\ &= E[(T_n - E[T_n])^2] + 2E[(T_n - E[T_n])(E[T_n] - \theta)] + E[(E[T_n] - \theta)^2] \\ &= \text{Var}(T_n) + 2 \times 0 + (E[T_n] - \theta)^2 \\ &= \text{variance de l'estimateur} + \text{biais}^2 \text{ de l'estimateur.} \end{aligned}$$

⇒ Ainsi, si T_n est un estimateur sans biais, le risque quadratique = $\text{Var}(T_n)$.

→ On a intérêt à ce qu'un estimateur soit sans biais et de faible variance !

- Rq:
- De 2 estimateurs sans biais, le meilleur est celui de + petite variance.
 - La variance de l'estimateur mesure sa variabilité : confronté à plusieurs jeux de données similaires, un estimateur de faible variance donnera des estimations de θ proches les unes des autres. Pour estimer correctement θ , il ne faut donc pas que la variance de l'estimateur soit trop grande.

(7)

→ On s'attend à ce que plus la taille des données augmente, plus l'estimation soit bonne (on a plus d'information).

Avec une observation infinie, on devrait même pouvoir estimer θ sans erreur!

On s'attend donc à ce que le risque de l'estimateur tende asymptotiquement vers 0, autrement dit que l'estimateur T_n converge (en un certain sens) vers θ .

Définition : L'estimateur T_n converge en moyenne quadratique vers θ si et seulement si son erreur quadratique moyenne tend vers 0 quand n tend vers l'infini : $T_n \xrightarrow{MQ} \theta \Leftrightarrow \lim_{n \rightarrow \infty} E[(T_n - \theta)^2] = 0$.

- Rq :
- Si T_n est sans biais, T_n CV en MQ $\Leftrightarrow \text{Var}(T_n) \xrightarrow{n \rightarrow \infty} 0$.
 - Le mieux estimateur possible de θ est un estimateur sans biais et de variance minimale (ESBVM). Il n'existe pas forcément.
 - Les notions d'exhaustivité et de compléter d'une statistique peuvent permettre de déterminer directement un ESBVM dans certains cas.
 - Dans ce qui suit, on utilise l'information de Fisher pour montrer parfois qu'un estimateur est un ESBVM.
 - A: rien ne dit que si T_n est un bon estimateur de θ , alors $\Psi(T_n)$ est un bon estimateur de $\Psi(\theta)$. En effet, on a souvent $E[T_n] = \theta$ et $E[\Psi(T_n)] \neq \Psi(\theta)$.
 Exemple → $f(x) = \ln(3 - x/3)$
 - b) Efficacité d'un estimateur et quantité d'information

La quantité d'information est un outil précieux pour évaluer la qualité d'un estimateur. Elle n'est définie que sous certaines conditions de régularité.

Définition : Pour $\theta \in \mathbb{R}$, si la loi des observations vérifie les conditions de régularité, on appelle quantité d'information (de Fisher) sur θ apportée par l'échantillon x_1, \dots, x_n la quantité :

$$I_n(\theta) = \text{Var}\left(\frac{\partial}{\partial \theta} \ln L(\theta; x_1, \dots, x_n)\right)$$

→ Sachant que $E\left[\frac{\partial}{\partial \theta} \ln L(\theta; x_1, \dots, x_n)\right] = 0$, on peut aussi écrire :

$$I_n(\theta) = E\left[\left(\frac{\partial}{\partial \theta} \ln L(\theta; x_1, \dots, x_n)\right)^2\right].$$

→ On peut aussi montrer l'écriture suivante, souvent utile dans les calculs :

$$I_n(\theta) = -E\left[\frac{\partial^2}{\partial \theta^2} \ln L(\theta; x_1, \dots, x_n)\right].$$

→ L'intérêt de la quantité d'information réside dans le fait qu'elle donne une borne inférieure pour la variance de n'importe quel estimateur sans biais !

Propriété : Inégalité de Fréchet-Darmois-Cramér-Rao (FDCR) : Si la loi des observations vérifie les conditions de régularité, alors pour tout estimateur T_n de θ , on a $\text{Var}(T_n) \geq \frac{\left(\frac{\partial}{\partial \theta} E[T_n]\right)^2}{I_n(\theta)}$

- Ainsi, si T_n est sans biais, on obtient $\text{Var}(T_n) \geq \frac{1}{I_n(\theta)}$ (borne de Cramér-Rao).
- La variance de n'importe quel estimateur sans biais de θ est donc forcément \geq à cette borne.

Definition : L'efficacité d'un estimateur est la quantité

$$\text{Eff}(T_n) = \frac{\left(\frac{\partial}{\partial \theta} E[T_n]\right)^2}{I_n(\theta) \text{Var}(T_n)}.$$

On a alors $0 \leq \text{Eff}(T_n) \leq 1$.

On dit que T_n est un estimateur efficace $\Leftrightarrow \text{Eff}(T_n) = 1$.

On dit que T_n est asymptotiquement efficace $\Leftrightarrow \lim_{n \rightarrow \infty} \text{Eff}(T_n) = 1$.

→ Rq :

- Si T_n est sans biais, $\text{Eff}(T_n) = \frac{1}{I_n(\theta) \text{Var}(T_n)}$
- Alors, si un estimateur sans biais est aussi efficace, alors sa variance est égale à la borne de Cramér-Rao \Rightarrow c'est un ESBVM.
- Parfois, il n'existe pas d'estimateur efficace. Alors s'il existe un ESBVM, sa variance est $>$ borne de Cramér-Rao.
- Si cette borne est grande, il est impossible d'estimer correctement θ !

⇒ Pour un estimateur sans biais, on peut donc juger de sa qualité en calculant son efficacité.

→ Quand les v.a. observées sont iid, on peut facilement voir que

$$I_n(\theta) = n I_1(\theta).$$

$$\begin{aligned} \text{En effet, } I_n(\theta) &= \text{Var}\left(\frac{\partial}{\partial \theta} \ln L(\theta; x_1, \dots, x_n)\right) = \text{Var}\left(\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(x_i; \theta)\right) \\ &= \text{Var}\left(\frac{\partial}{\partial \theta} \sum_{i=1}^n \ln f(x_i; \theta)\right) = \text{Var}\left(\sum_{i=1}^n \frac{\partial}{\partial \theta} \ln f(x_i; \theta)\right) = \sum_{i=1}^n \text{Var}\left(\frac{\partial}{\partial \theta} \ln f(x_i; \theta)\right) \\ &= n I_1(\theta). \end{aligned}$$

Exercice - Auton Chap 6 → ex. 5

④ - Autres notions et quelques extensions

On revient dans ce paragraphe sur la vraisemblance et l'information.

→ Si les données étaient dépendantes, il faudrait adopter la formule de la vraisemblance. En notant f toutes les densités des v.a. X_i , on aurait

$$L(\theta; \mathbf{x}) = f(\mathbf{x}; \theta) = f(x_1; \theta) \prod_{i=2}^n f(x_i; \theta; x_1, \dots, x_{i-1}).$$

→ La vraisemblance est invariante par transformation bijective du jeu de données.

Si $z = \Psi^{-1}(\mathbf{x})$ où Ψ est un difféomorphisme, alors les densités vérifient

$$f_z(z; \theta) = f_x(\Psi(z); \theta) \left| \frac{d\Psi}{dz}(z) \right|, \text{ avec le Jacobien qui est clairement une constante multiplicative.}$$

Définition : lorsque $\max_{\theta'} L(\theta') < +\infty$, pour fixer la constante multiplicative qui apparaît parfois dans la définition de la vraisemblance, on définit la vraisemblance relative par $RL(\theta) = \frac{L(\theta)}{\max_{\theta'} L(\theta')}$

Rq: Dans l'option où l'on ne pourra pas chercher $\hat{\theta}_n$ par optimisation, on peut considérer comme "plausibles" toutes les valeurs de θ qui ont une vraisemblance relative suffisamment grande (donc pour les θ t.q. $RL(\theta) > c$, avec c un seuil choisi).

→ $\hat{\theta}_n$ est une fonction de X , $\hat{\theta}_n = \hat{\theta}_n(X)$: en effet, autre jeu de données conduirait à une estimation $\hat{\theta}_n$ différente. (On conserve le même estimateur, mais l'estimation est différente).

→ Approximation quadratique: on considère que l'estimateur du max. de vraisemblance se réalise à un endroit unique. En notant $\hat{\theta}_n = \hat{\theta}_n(X)$, et en faisant un développement de Taylor à l'ordre 2 au voisinage de $\hat{\theta}_n$, on obtient: en notant $f(\theta) = \ln L(\theta)$ et en supposant que f est C^2 , alors

$$\forall \theta \in \text{v}(\hat{\theta}), f(\theta) = f(\hat{\theta}) + \frac{(\theta - \hat{\theta})}{1!} f'(\hat{\theta}) + \frac{(\theta - \hat{\theta})^2}{2!} f''(\hat{\theta}) + o(\|\theta - \hat{\theta}\|^2)$$

Ainsi, $\ln L(\theta) = \ln L(\hat{\theta}) + (\theta - \hat{\theta}) \underbrace{(\ln L(\theta))'}_{=0 \text{ par déf. de } \hat{\theta}} + \frac{1}{2} (\theta - \hat{\theta})^2 (\ln L(\hat{\theta}))'' + o(\|\theta - \hat{\theta}\|^2)$

D'où

$$\ln \left(\frac{L(\theta)}{L(\hat{\theta})} \right) = \frac{1}{2} (\theta - \hat{\theta}) \frac{\frac{d^2}{d\theta^2} \ln L(\theta)}{|_{\theta=\hat{\theta}}} + o(\|\theta - \hat{\theta}\|^2), \text{ soit}$$

sous forme matricielle: $\ln \left(RL(\theta) \right) = \frac{1}{2} (\theta - \hat{\theta})^T \underbrace{\left[\frac{d^2}{d\theta^2} \ln L(\theta) \right]_{|_{\theta=\hat{\theta}}}}_{\text{matrice hessienne prise en } \hat{\theta}} (\theta - \hat{\theta}) + o(\|\theta - \hat{\theta}\|^2)$

En prenant $\theta = \hat{\theta} + \frac{n}{\sqrt{n}}$ avec n fixé (on est bien au voisinage de $\hat{\theta}$),

on obtient $\ln RL(\theta) = \ln RL\left(\hat{\theta} + \frac{n}{\sqrt{n}}\right) = \frac{1}{2} \left(\frac{1}{\sqrt{n}} \right)^2 \left(n^T \left[\frac{d^2}{d\theta^2} \ln L(\theta) \right]_{|_{\theta=\hat{\theta}}} n \right) +$

→ Quand $n \uparrow$, l'ensemble des θ plausibles \downarrow : en effet, le $o\left(\left(\frac{1}{\sqrt{n}}\right)^2\right)$ term de droite \downarrow plus vite et la log-vraisemblance relative excède moins souvent le seuil... C'est la même intuition que le rapprochement de l'IC grand on a plus d'observations - (if + loin).

→ Notion de score: pour chercher le maximum de vraisemblance, on cherche des points qui annulent le gradient de la log-vraisemblance:

Définition: le score, noté U , est défini comme $U(\theta) = \frac{d}{d\theta} \ln L(\theta)$. On le note aussi $U(\theta; X) = \frac{d}{d\theta} \ln L(\theta; X)$. On a $E_{\theta}[U(\theta; X)] = 0$ quand $X \sim P_{\theta}$.

→ Notion d'exhaustivité d'une statistique : (voir aussi polycop. D. Gardoin - Stat. Inf. Ancien. § 2.3, 2.4, 2.5) On considère un modèle statistique paramétrique $(X, \mathcal{A}, \{P_\theta; \theta \in \Theta \subset \mathbb{C}^{d_\theta}\})$, et on essaie d'obtenir le plus de connaissance possible sur θ à partir de $x \in X$ observé. Sachant que x est un vecteur de grande dimension (n), il peut être intéressant de résumer les données en une stat. t(x) de dimension $\ll n$. Notamment, t(x) contientre θ d'infos que x sur θ , mais une statistique exhaustive conserve l'intégralité de l'information sur θ .

Définition : Une statistique t est exhaustive pour $\theta \Leftrightarrow$ la loi de probabilité conditionnelle de X sachant $T=t$ ne dépend pas de θ .

Interprétation : si la loi de X sachant $(T=t)$ ne dépend pas de θ , cela veut dire que le seul résumé t(x) de l'observation suffit pour estimer θ car la connaissance de x n'y apporte aucune info supplémentaire sur θ . Par conséquent, on peut ne se servir que de t(x) pour estimer θ .

Exemple : Soit le modèle $(\{0,1\}, \mathcal{P}(\{0,1\}), \{B(p); p \in [0,1]\})^n$ des pièces defectueuses. Intuitivement, connaître la proportion de pièces defectueuses dans l'échantillon suffit à comme connaître l'information pour estimer p . On s'attend donc à ce que $p_n(x) = \frac{1}{n} \sum_{i=1}^n x_i$ soit une statistique exhaustive, ou bien même encore $t(x) = \sum_{i=1}^n x_i$. Vérifions !

On sait que $T = \sum_{i=1}^n X_i \sim B(n, p)$.

$$\begin{aligned} \text{IP}(X=x | T=t) &= \text{IP}(X_1=x_1, \dots, X_n=x_n | \sum_{i=1}^n X_i = t) = \frac{\text{IP}(X_1=x_1, \dots, X_n=x_n, \sum X_i=t)}{\text{IP}(\sum_{i=1}^n X_i = t)} \\ &= \begin{cases} 0 & \text{si } \sum x_i \neq t \\ \text{IP}(X_1=x_1, \dots, X_n=x_n) / \text{IP}(\sum X_i=t) & \text{si } \sum x_i = t. \end{cases} \quad i \text{ avec } \text{IP}(X_i=x_i) = p^{x_i} (1-p)^{1-x_i}. \end{aligned}$$

Comme les X_i sont $\perp \!\! \perp$, $\frac{\text{IP}(X_1=x_1, \dots, X_n=x_n)}{\text{IP}(\sum X_i=t)} = \frac{\prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}}{C_n^t p^t (1-p)^{n-t}} = \frac{p^{\sum x_i} (1-p)^{n-\sum x_i}}{C_n^t p^t (1-p)^{n-t}} = \frac{1}{C_n^t}$

$$\text{Ainsi } P(X=x \mid T=t) = \begin{cases} 0 & \text{si } \sum_{i=1}^n x_i \neq t, \\ \frac{1}{C_n^t} & \text{si } \sum_{i=1}^n x_i = t. \end{cases}$$
(10)

\Rightarrow Cette loi (uniforme) ne dépend pas de p , donc $t(x) = \sum x_i$ est exhaustive pour θ .

Remarque: On voit bien à travers cet exemple qu'il n'est pas toujours facile de vérifier cette propriété... On utilise alors le résultat suivant caractérisant l'exhaustivité.

Théorème de factorisation de Fisher-Neyman:

t est exhaustive pour $\theta \Leftrightarrow$ il existe 2 fonctions mesurables g et h telles que:

$$\forall x \in X, \forall \theta \in \Theta, L(\theta; x) = g(t(x); \theta) h(x).$$

Preuve: Rely sur O. Gaudoin.

Exemple: En reprenant l'exemple précédent, $L(p; x_1, \dots, x_n) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}$

Cette vraisemblance est de la forme: $g(\sum x_i; p) \cdot h(x_{-1}) = p^{\sum x_i} (1-p)^{n - \sum x_i}$

Définition: Une statistique exhaustive $s(x)$ est dite minimale si, quelle que soit la statistique exhaustive $t(x)$, on peut écrire t comme une fonction de s .

En pratique, on obtient une stat. exhaustive minimale en regardant la vraisemblance à une constante multiplicative près, et après simplification, en regardant quelle fonction des données est nécessaire et suffisante pour la calculer.

Exercice \rightarrow Ainsi Chap 6 ex 3

(5)- Propriétés des EMV et des EMV.

a) le cas des Estimateurs par la méthode des moments (EMM):

→ On a vu que si $\theta = \mathbb{E}(X)$, alors l'EMM de θ est $\tilde{\theta}_n = \bar{X}_n$.

Cette méthode est justifiée par la loi des grands nombres : $\bar{X}_n \xrightarrow[n \rightarrow \infty]{\text{P.s.}} \mathbb{E}(X)$.

Donc si $\theta = \mathbb{E}(X)$, \bar{X}_n est un estimateur convergent presque sûrement (asymptotique).

→ Par ailleurs, on peut aussi constater que \bar{X}_n est un bon estimateur de $\theta = \mathbb{E}(X)$, sans utiliser la loi des grands nombres, car :

- il est sans biais : $E(\bar{X}_n) = \theta$
 - $\text{Var}(\bar{X}_n) = \frac{\text{Var}(X)}{n} \xrightarrow[n \rightarrow \infty]{\text{P.s.}} 0$
- $\Rightarrow \bar{X}_n$ est un estimateur sans biais et
convergent en moyenne quadratique de $\mathbb{E}(X)$.

→ Cas de la variance: en prenant comme estimateur de $\text{Var}(X)$ la quantité :

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2.$$

- biais: $E(S_n^2) = E\left[\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2\right] = \frac{1}{n} \sum_{i=1}^n E(X_i^2) - E(\bar{X}_n^2) = E(X^2) - E(\bar{X}_n^2)$,
 $= \text{Var}(X) + \mathbb{E}(X)^2 - \text{Var}(\bar{X}_n) - E(\bar{X}_n)^2 = \text{Var}(X) + \mathbb{E}(X)^2 - \frac{\text{Var}(X)}{n} - \mathbb{E}(X)^2$
 $= \left(1 - \frac{1}{n}\right) \text{Var}(X) \Rightarrow$ biaisé ! Mais $S_n'^2 = \frac{n}{n-1} S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ est sans biais
- $\text{Var}(S_n'^2) = \frac{1}{n(n-1)} \left[(n-1) E[(X - \mathbb{E}(X))^4] - (n-3) \text{Var}(X)^2 \right] \xrightarrow[n \rightarrow \infty]{\text{P.s.}} 0$.

→ $S_n'^2$ est sans biais et convergent en moyenne quadratique !

Rq: en R, c'est bien $S_n'^2$ que l'on obtient (et pas S_n^2). (par la commande var(x))

- Aucun résultat sur la qualité de S_n comme estimateur de l'écart-type de la loi de X .
- Un EMM est asymptotiquement sans biais et convergent presque sûrement.
- $\text{Cov}(\bar{X}_n, S_n'^2) = \frac{1}{n} E[(X - \mathbb{E}(X))^3] \Rightarrow \bar{X}_n$ et $S_n'^2$ sont corrélés... mais asymptotiquement non-correlés. Seul si la loi de X_i est gaussienne conduit à l'indépendance de ces 2 estimateurs.

b) Le cas des EMV.

(11)

Un estimateur de maximum de vraisemblance n'est pas forcément unique (il peut y avoir plusieurs maxima), ni sans biais, ni de variance minimale, ni efficace. Mais il a de très belles propriétés asymptotiques, à condition que les hypothèses suivantes soient respectées :

- Hyp:
 - les données proviennent d'un modèle iid (n observations);
 - Θ est compact, de dimension d ;
 - $\theta^* \in \Theta$ (la vraie valeur) est dans l'intérieur de Θ (utile pour Taylor)
 - le modèle est identifiable (si $\theta \neq \theta'$, alors P_θ et $P_{\theta'}$ sont 2 lois distinctes).
 - il existe un voisinage $v(\theta^*)$ dans Θ avec :
 - la log-vraisemblance est C^3 sur $v(\theta^*)$
 - $\forall r, s, t, \theta \mapsto \frac{1}{n} \mathbb{E}_\theta \left[\frac{\partial^3}{\partial \theta_r \partial \theta_s \partial \theta_t} \ln L(\theta; X) \right]$ uniformément borné
 - $\forall \theta \in v(\theta^*)$, $I(\theta)$ est une matrice définie positive, et $\text{Var}_{\theta^*} I(\theta)$.
 - $\forall r, s, I_{rs}(\theta) = \mathbb{E}_\theta \left[\frac{\partial}{\partial \theta_r} \ln L(\theta; X) \frac{\partial}{\partial \theta_s} \ln L(\theta; X) \right]$.

Sous ces conditions, on a :

- Propriété : si les X_i sont iid, dépendant d'un paramètre θ réel, la loi des X_i vérifiant les conditions de régularité ci-dessus, on a :
- $\hat{\theta}_n$ converge presque sûrement vers θ^* ; ($\mathbb{P}_{\theta^*}(\hat{\theta}_n \rightarrow \theta^*) = 1$)
 - $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{\mathcal{D}} N(0, 1)$; quand n est grand, $\hat{\theta}_n \sim \mathcal{N}(\theta^*, \frac{1}{I(\theta^*)})$.
 $\Rightarrow \hat{\theta}_n$ est asymptotiquement sans biais et efficace.
 - En général l'EMV est meilleur que l'GMM au sens où $\text{Var}(\hat{\theta}_n) \leq \text{Var}(\tilde{\theta}_n)$. (au moins vrai asymptotiquement).

→ [Méthode Delta]: Si $\hat{\theta}_n$ est l'EMV de θ , alors $\psi(\hat{\theta}_n)$ est l'EMV de $\psi(\theta)$. De plus, si ψ est dérivable, on a:

$$V_n \left(\psi(\hat{\theta}_n) - \psi(\theta) \right) \xrightarrow[n \rightarrow \infty]{\mathcal{D}} N\left(0, \frac{\psi'(\theta)^2}{I_n(\theta)}\right)$$

ou encore

$$\boxed{\left[\psi(\hat{\theta}_n) \sim N\left(\psi(\theta), \frac{\psi'(\theta)^2}{I_n(\theta)}\right) \right]}.$$

→ Exemple: loi de Bernoulli: $X_i \sim B(p)$

- EMM de p : $\hat{p}_n = \bar{X}_n$
- EMV de p : $\hat{p}_n = \bar{X}_n$.

→ \bar{X}_n est un estimateur sans biais de p , $E[X] = p$. Or $E[X_i] = p$, donc \hat{p}_n est un estimateur sans biais de p .

→ De plus, $Var(\bar{X}_n) = \frac{Var(X)}{n} = \frac{1}{n} p(1-p) \xrightarrow[n \rightarrow \infty]{} 0$, donc \hat{p}_n CV en moy. quadratique.

→ Enfin, $I_n(p) = Var\left(\frac{d}{dp} \ln L(p; X_1, \dots, X_n)\right) = Var\left[\frac{\sum_{i=1}^n X_i - np}{p(1-p)}\right] = \frac{Var(\sum_{i=1}^n X_i)}{p^2(1-p)^2}$
 $\sum_{i=1}^n X_i \sim B(n, p)$

$$\Downarrow \frac{n p(1-p)}{p^2(1-p)^2} = \frac{n}{p(1-p)} \Rightarrow Var(\hat{p}_n) = \frac{1}{I_n(p)} \Rightarrow \hat{p}_n \text{ est efficace.}$$

⇒ \hat{p}_n est un ESBVR de p .

→ On a vu que si θ est le paramètre recherché, on a $\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{\mathcal{D}} N\left(\theta, \frac{1}{I_n(\theta)}\right)$.
 En pratique on ne connaît pas θ (on cherche à l'estimer), on approxime donc $I_n(\theta)$ par $I_n(\hat{\theta})$.

III - Estimation par intervalle (ensemble) - INTERVALLE DE CONFIANCE

(12)

① - Problématique

Dans la partie précédente, on a estimé θ par une unique valeur ($\tilde{\theta}_n$ ou $\hat{\theta}_n$). C'était une estimation ponctuelle. On peut s'attendre à ce que cette estimation soit proche de θ , mais il y a peu de chance (voire aucune dans le cas continu) qu'elle soit parfaite. En effet, si la loi de $\hat{\theta}_n$ est continue, on sait que $P(\hat{\theta}_n = \theta) = 0$.

C'est donc naturel d'estimer θ en proposant un ensemble de valeurs raisonnables et vraisemblables pour θ , proches de $\hat{\theta}_n$. C'est une région de confiance.

On suppose dans la suite que $\theta \in \mathbb{R}$: la région sera donc un intervalle.

Définition : Un intervalle de confiance de seuil $\alpha \in [0,1]$ pour un paramètre θ est un intervalle aléatoire I tel que $P(\theta \in I) = 1-\alpha$.

- Rq : • α est une probabilité d'erreur, donc si possible fixé petit.
• I est ALÉATOIRE : ses bornes sont aléatoires, car elles s'expriment par des fonctions des X_i . Dans l'événement $P(\theta \in I)$, θ est déterministe (et inconnu), et $I = [z_1, z_2]$ sont des (bornes) aléatoires. Si on note z_1 et z_2 les réalisations de ces bornes pour une expérience donnée, il est faux de dire " " θ à 95% de chance de se trouver entre z_1 et z_2 " ". Par contre, il est correct de dire : " " θ à 95% de chance de se trouver entre z_1 et z_2 " ". En effet, θ est ou n'est pas dans l'intervalle $[z_1, z_2]$ dont la probabilité est 0 ou 1. Mais si on répète 100 fois l'expérience, en moyenne θ sera 95 fois dans l'intervalle $[z_1, z_2]$.

②. Procédé

Il apparaît logique de proposer comme intervalle de confiance un ensemble de valeurs centré sur l'estimateur EMV pertinent $\hat{\theta}_n$. Donc I sera de la forme $I = [\hat{\theta}_n - \varepsilon, \hat{\theta}_n + \varepsilon]$. Ensuite, il reste à déterminer ε tel que :

$$P(\theta \in I) = P(\hat{\theta}_n - \varepsilon \leq \theta \leq \hat{\theta}_n + \varepsilon) = P(|\hat{\theta}_n - \theta| \leq \varepsilon) = 1 - \alpha.$$

Parfois, cette démarche n'aboutit pas, car ε et α ne doivent pas dépendre de θ pour que I soit utilisable. En fait, on ne peut déterminer un tel ε que si la loi de proba de $\hat{\theta}_n - \theta$ ne dépend pas de θ .

Plus globalement, la technique la plus efficace pour trouver un intervalle de confiance (IC) consiste donc à chercher une fonction pivotale (ou un pivot), c'est à dire une v.a. fonction de θ et des observations X_1, \dots, X_n ; mais dont la loi ne dépend pas de θ !

→ Interprétation d'un IC: si l'IC est petit, l'ensemble des valeurs raissemblables pour θ est resserré autour de $\hat{\theta}_n$. Sinon c'est l'inverse. Ainsi, un IC construit à partir d'un estimateur permet d'appréhender la précision de cet estimateur.

③. Exemples de construction

a) Paramétriser de la loi normale:

Si X_1, \dots, X_n sont iid, avec $X_i \sim N(\mu, \sigma^2)$, on sait que :

- L'ESBM de μ est \bar{X}_n .
- S_n^2 est un estimateur asymptotiquement sans biais, qui converge en moyenne quadratique.

→ IC pour la moyenne m :

(13)

. la première idée est de chercher un IL pour m de la forme $[\bar{X}_n - \varepsilon; \bar{X}_n + \varepsilon]$.

Donc le problème revient pour α fixé à chercher ε tel que:

$$P(|\bar{X}_n - m| \leq \varepsilon) = 1 - \alpha.$$

- On sait que $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$, et que $\sum_{i=1}^n X_i \sim \mathcal{N}(nm, n\sigma^2)$. Donc on a $\bar{X}_n \sim \mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$. Ainsi, la.v.a. $V = \frac{\bar{X}_n - m}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0, 1)$ est une fonction pivotale!
- On obtient donc $V = \frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$, et

$$P(|\bar{X}_n - m| \leq \varepsilon) = P\left(|V| \leq \frac{\varepsilon}{\sigma/\sqrt{n}}\right) = 1 - P\left(|V| > \frac{\varepsilon}{\sigma/\sqrt{n}}\right) = 1 - \alpha.$$

Ainsi, $u_\alpha = \frac{\varepsilon}{\sigma/\sqrt{n}}$ et donc $\varepsilon = \frac{\sigma}{\sqrt{n}} u_\alpha$, où u_α est le quantile adéquat.

Propriété: Un IC de seuil α pour la moyenne m de la loi $\mathcal{N}(m, \sigma^2)$ est

$$IC_\alpha(m) = \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} u_\alpha; \bar{X}_n + \frac{\sigma}{\sqrt{n}} u_\alpha \right] - \underbrace{Q_{1-\frac{\alpha}{2}}}_{\text{car IC symétrique}}$$

. Pb: ce n'est utilisable que si σ est connu... On peut alors remplacer σ par un estimateur, par exemple S_n' . Mais $\frac{\bar{X}_n - m}{S_n'}$ n'est pas $\mathcal{N}(0, 1)$!
Donc on n'a pas d'IL pour m en faisant ceu. S_n'

. Solution: on résoud le problème en remarquant que $\frac{\bar{X}_n - m}{S_n'} = \sqrt{n-1} \frac{\bar{X}_n - m}{S_n}$ suit une loi de Student $St(n-1)$. Et on aboutit aujour à :

Propriété: Un IL de seuil α pour la moyenne m de la loi $\mathcal{N}(m, \sigma^2)$ est

$$IC_\alpha(m) = \left[\bar{X}_n - \frac{S_n'}{\sqrt{n}} t_{n-1, \alpha}; \bar{X}_n + \frac{S_n'}{\sqrt{n}} t_{n-1, \alpha} \right] -$$

→ Exercice et interprétation:

$t_{n-1, \alpha}$ car (IC bilatéral
loi de Student symétrique)

- Remarques: on peut considérer des IC unilatéraux. Ici l'IC était bilatéral... mais pour intervalle I tel que $P(m \in I) = 1-\alpha$ convient.

Par exemple, $IC_{\alpha}(m) = [\bar{X}_n - \frac{s_n}{\sqrt{n}} t_{n-1, 2}, +\infty)$; $t_{n-1, 2} = \frac{s_n}{\sqrt{n}} q_{1-\alpha}$ fournit une borne inférieure pour l'estimation de m au seuil α .

→ IC pour la variance σ^2 :

- On rappelle que S_n^2 , estimateur de σ^2 , est donné par

$$S_n^2 = \frac{1}{n} \sum (x_i - \bar{x}_n)^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x}_n)^2$$

- D'après le Théorème de Fisher, $\frac{n S_n^2}{\sigma^2}$ est de loi χ_{n-1}^2 .

Alors, $\forall (a, b) \in \mathbb{R}^2$ t.p. $0 < a < b$,

$$P(a \leq \frac{n S_n^2}{\sigma^2} \leq b) = P\left(\frac{n S_n^2}{b} \leq \sigma^2 \leq \frac{n S_n^2}{a}\right) \text{ d'une part}$$

$$= F_{\chi_{n-1}^2}(b) - F_{\chi_{n-1}^2}(a) \text{ d'autre part.}$$

- Il y a une infinité de possibilités de choisir a et b pour que cette proba égale $1-\alpha$. En général, on "équilibre" les risques, c.-à-d on prend a et b t.p.

$$F_{\chi_{n-1}^2}(b) = 1 - \frac{\alpha}{2} \quad \text{et} \quad F_{\chi_{n-1}^2}(a) = \frac{\alpha}{2} -$$

- Alors, en notant z les quantiles de la loi du χ^2 , on obtient:

Propriété: Un IL de seuil α pour la variance σ^2 d'une loi $\mathcal{N}(m, \sigma^2)$ est:

$$IC_{\alpha}(\sigma^2) = \left[\frac{n S_n^2}{\chi_{n-1, \frac{\alpha}{2}}}, \frac{n S_n^2}{\chi_{n-1, 1-\frac{\alpha}{2}}} \right]$$

b) Cas d'une proportion :

Il s'agit ici de déterminer l'IC pour le paramètre p de la loi de Bernoulli, en disposant d'un échantillon X_1, \dots, X_n de cette loi.

Nous avons déjà vu que l'ESTVM de p vaut $\hat{p}_n = \bar{X}_n$.

En reprenant l'exemple des pièces defectueuses, admettons que sur 800 pièces testées, 420 sont defectueuses. Les pièces sont \mathbb{I} , et la proba qu'une pièce prise au hasard soit defectueuse vaut p . On cherche à proposer une estimation ensemble pour p . Ainsi, $X_i \sim \mathcal{B}(p)$, et on ne connaît pas le détail de toutes les pièces defectueuses, mais on sait par contre que $\hat{p}_n = \frac{420}{800} = 52,5\%$.

On s'intéresse à un IC de seuil 5% pour p , autrement dit on cherche I_k .

$$\mathbb{P}(p \in I) = 1-\alpha = 95\%.$$

Dans ce cas, il n'est pas facile de trouver une fonction pivotale... En effet, en considérant $\sum X_i \sim \mathcal{B}(n, p)$, on a du mal à monopoler ensuite la Binomiale...

Il existe un IC exact basé sur la loi de Fisher-Snedecor, mais on utilise plus communément l'approximation de la loi Binomiale par la loi Normale.

En effet

$$\frac{\sum X_i - nm}{\sqrt{np^2}} = \frac{\frac{1}{n} \sum X_i - m}{\frac{\sigma}{\sqrt{n}}} = \sqrt{n} \frac{\bar{X}_n - E(X)}{\text{Var}(X)} \underset{\substack{\uparrow \\ \text{approx.}}}{\sim} \mathcal{N}(0, 1).$$

Ainsi, $\sqrt{n} \frac{\hat{p}_n - p}{\sqrt{p(1-p)}}$ est approximativement de loi $\mathcal{N}(0, 1)$.

On a donc la fonction pivotale : en notant $T = \sum_{i=1}^n X_i$, on a $\mathbb{P}\left(\left|\frac{T-np}{\sqrt{np(1-p)}}\right| \leq u_\alpha\right) \approx 1-\alpha$.

Donc on cherche à écrire $\left|\frac{T-np}{\sqrt{np(1-p)}}\right|$ sous la forme $Z_1 \leq p \leq Z_2$.

$$\left| \frac{T-np}{\sqrt{np(1-p)}} \right| \leq u_\alpha \Leftrightarrow \frac{(T-np)^2}{np(1-p)} \leq u_\alpha^2 \Leftrightarrow p^2(n+u_\alpha^2) - p(2T+u_\alpha^2) + \frac{T^2}{n} \leq 0.$$

Le discriminant en p est toujours positif, sauf entre ses racines, donc ses 2 racines sont les bornes de l'intervalle recherché. On obtient

$$\left[\frac{T + \frac{u_\alpha^2}{2n} - u_\alpha \sqrt{\frac{u_\alpha^2}{4n^2} + \frac{T(n-T)}{n^3}}}{1 + \frac{u_\alpha^2}{n}}, \frac{T + \frac{u_\alpha^2}{2n} + u_\alpha \sqrt{\frac{u_\alpha^2}{4n^2} + \frac{T(n-T)}{n^3}}}{1 + \frac{u_\alpha^2}{n}} \right].$$

En remarquant que $\hat{p}_n = \frac{T}{n}$, en négligeant u_α^2 par rapport à n (vrai pour des valeurs courantes de α et n), on obtient un IC asymptotique:

Propriété: Un IC asymptotique de seuil α pour le paramètre p vaut:

$$IC_{1-\alpha}(p) = \left[\hat{p}_n - u_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + u_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \right]$$

$q_{1-\frac{\alpha}{2}}^{U(0,1)}$

Application: dans l'exemple des pièces defectueuses, on obtient un IC asymptotique

$$IC(p) \approx [0,49; 0,56].$$

On a donc 95% confiance dans le fait que p (le % de pièces defectueuses) sera compris dans cet intervalle. Cependant, on voit qu'il est possible que $p < 50\%$, auquel cas le nb de pièces defectueuses n'est pas majoritaire.

Rq: On voit bien qu'on peut élargir ou retrécir cet intervalle de confiance en jouant sur 2 facteurs principalement: n et α .

Exercice Annexe chap 4 → 14, 15, 16

IV - Ouverture vers les Tests

(15)

On profite ici des propriétés de la méthode du maximum de vraisemblance pour présenter deux types de statistique de test.

① Test de Wald :

On dispose ici d'un estimateur $\hat{\theta}_n$ de θ asymptotiquement normal.

Ainsi, $\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \xrightarrow[n \rightarrow \infty]{\mathcal{D}} N(0,1)$, où $\hat{\sigma}_n$ est un estimateur de l'écart-type de $\hat{\theta}_n$.

Supposons que l'on veuille tester

$H_0: \theta \leq \theta_0$ contre $H_1: \theta > \theta_0$ avec θ_0 valeur fixe.

On peut prendre une règle de décision du type:

- si $\frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n} > t$, alors on rejette H_0 , ($t > 0$)
- sinon on ne rejette pas H_0 .

On peut choisir t tel que

$$\begin{aligned} \alpha &= P_{(\theta_0)}(\text{rejeter } H_0) = \sup_{\theta \leq \theta_0} P_\theta \left(\frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n} > t \right) = \sup_{\theta \leq \theta_0} P_\theta \left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} + \frac{\theta - \theta_0}{\hat{\sigma}_n} > t \right) \\ &\stackrel{\text{erreur de premier espèce (niveau)}}{\approx} \sup_{\theta \leq \theta_0} P_\theta \left(Z + \frac{\theta - \theta_0}{\hat{\sigma}_n} > t \right) \text{ avec } Z \sim N(0,1). \end{aligned}$$

La fonction $\theta \mapsto P_\theta \left(Z + \frac{\theta - \theta_0}{\hat{\sigma}_n} > t \right)$ est 1 en θ , donc le sup est atteint en θ_0 .

Ainsi $\alpha \approx P(Z > t)$, et t est donc le quantile d'ordre $1-\alpha$ de la loi $N(0,1)$.

② - Rapport de vraisemblance

→ On rappelle qu'on note d la dimension de Θ . Si les hypothèses de régularité décrites avant sont satisfaites, on définit alors la statistique du rapport de vraisemblance, $W(\theta)$ par :

$$W(\theta) = -2 \ln RL(\theta) = 2 (\ln L(\hat{\theta}) - \ln L(\theta)) , \text{ ou bien}$$

$$W(\theta; x) = -2 \ln RL(\theta; x) = 2 (\ln L(\hat{\theta}; x) - \ln L(\theta; x)) .$$

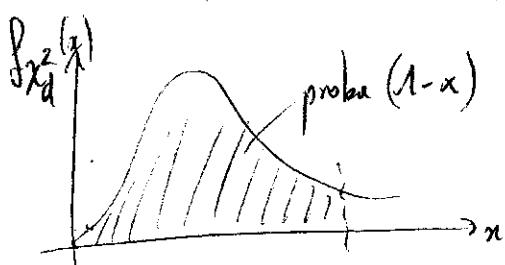
Au vu de l'approximation quadratique de $RL(\theta; x)$ vu précédemment par un développement de Taylor, on sait que lorsque $n \rightarrow +\infty$, on a

$$\begin{matrix} W(\theta_0; x) \\ \nearrow \end{matrix} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} X_d^2 \quad (\text{car somme de gaussiennes au carré})$$

vrai paramètre.

⇒ On peut donc utiliser ceu par avoir une région de confiance asymptotique sur θ_0 en utilisant $W(\theta_0; x)$ comme pivot:

$$P_{\theta_0} (W(\theta_0; x) \leq c_{1-\alpha}^d) \approx 1-\alpha , \text{ où } c_{1-\alpha}^d \text{ est le quantile d'ordre } 1-\alpha \text{ de la loi de } X_d^2 .$$



Rq: $W(\theta)$ petit si $\hat{\theta}$ proche de θ .

Comme $W(\theta) = 2 (\ln L(\hat{\theta}) - \ln L(\theta))$, on peut voir

$$\begin{aligned} \underbrace{W(\theta)}_{\geq 0} &\leq c_{1-\alpha}^d \Leftrightarrow \left\{ \theta \in \mathbb{R} : \ln L(\theta) \geq \ln L(\hat{\theta}) - \frac{1}{2} c_{1-\alpha}^d \right\} \text{ comme une} \\ &\equiv 2 \ln \left(\frac{L(\hat{\theta})}{L(\theta)} \right) \leq c_{1-\alpha}^d \quad \text{région de confiance de niveau asymptotique } (1-\alpha). \end{aligned}$$

→ En généralisant au test statistique, supposons que l'on veuille tester $H_0: \theta \in \Theta_0$ contre $H_1: \theta \in \Theta_1$, avec $\Theta_0 \cup \Theta_1$ partition de Θ . Il est logique et cohérent de rejeter H_0 si l'observation sur X est beaucoup plus probable lorsque θ varie dans Θ_1 , donc si :

$$\sup_{\theta \in \Theta_1} L(\theta; X) \gg \sup_{\theta \in \Theta_0} L(\theta; X).$$

On utilise alors la statistique $T(X) = 2 \ln \left(\frac{\sup_{\theta \in \Theta_1} L(\theta; X)}{\sup_{\theta \in \Theta_0} L(\theta; X)} \right)$.
Ainsi, T dépend de X mais pas de θ .

Une région de rejet naturelle est $R = \{X \text{ tel que } T(X) \geq t\}$ avec t à choisir en fonction du niveau du test.

Pour choisir t , on doit donc connaître la loi de $T(X)$ sous (H_0) (ou sa loi asymptotique).